Turner, R. M., Rhodes, K. M., Jones, H. E., Higgins, J. P. T., Haskins, J., Whiting, P. F., Hróbjartsson, A., Caldwell, D. M., Morris, R. W., Reeves, B. C., Worthington, H., Boutron, I., & Savović, J. (2020). Agreement was moderate between data-based and opinion-based predictions of biases affecting randomized trials within meta-analyses. *Journal of Clinical Epidemiology*, *125*, 16-25. https://doi.org/10.1016/j.jclinepi.2020.05.009

Peer reviewed version

Link to published version (if available):
10.1016/j.jclinepi.2020.05.009

Agreement was moderate between data-based and opinion-based assessments of biases affecting randomised trials within meta-analyses

Rebecca M Turner[1,2], Kirsty M Rhodes[2,3], Hayley E Jones[4], Julian PT Higgins[4], Jess Haskins[4], Penny Whiting[4], Asbjørn Hróbjartsson[5,6,7], Debbi Caldwell[4], Richard Morris[4], Barney C Reeves[8], Helen Worthington[9], Isabelle Boutron[10,11,12], Jelena Savović[4,13]

[1]MRC Clinical Trials Unit, University College London, London, UK
[2]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK
[3]Statistical Innovation, Oncology Biometrics, AstraZeneca, Cambridge, UK
[4]Population Health Sciences, Bristol Medical School, University of Bristol, UK
[5]Centre for Evidence-Based Medicine Odense (CEBMO), Odense University Hospital, Odense, Denmark
[6]Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[7]Open Patient data Explorative Network (OPEN), Odense University Hospital, Odense, Denmark
[8]Clinical Trials and Evaluation Unit, Bristol Trials Centre, Bristol Medical School, University of Bristol, Bristol, UK
[9]Division of Dentistry, School of Medical Sciences, University of Manchester, Manchester, UK
[10]Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Assistance Publique Hôpitaux de Paris, Paris, France
[11]Team METHODS, Centre of Research in Epidemiology and Statistics–CRESS Inserm UMR1153, Paris, France
[12]Université Paris Descartes, Paris, France
[13]NIHR Applied Research Collaboration (ARC) West, University Hospitals Bristol NHS Foundation Trust, Bristol, UK

Corresponding author: Rebecca Turner, MRC Clinical Trials Unit, University College London, London, UK. E-mail: becky.turner@ucl.ac.uk Tel: +44(0)20 7670 4644

Agreement was moderate between data-based and opinion-based assessments of biases affecting randomised trials within meta-analyses

ABSTRACT

Background: Randomised trials included in meta-analyses are often affected by bias caused by methodological flaws or limitations, but the degree of bias is unknown. Two proposed methods adjust trial results for bias using: (1) empirical evidence from published meta-epidemiological studies; or (2) expert opinion.

Methods: We investigated agreement between data-based and opinion-based approaches to assessing bias in each of four domains: sequence generation, allocation concealment, blinding and incomplete outcome data. From each sampled meta-analysis, a pair of trials with the highest and lowest empirical model-based bias estimates was selected. Independent assessors were asked which trial within each pair was judged more biased on the basis of detailed trial design summaries.

Results: Assessors judged trials to be equally biased in 68% of pairs evaluated. When assessors judged one trial as more biased, the proportion of judgements agreeing with the model-based ranking was highest for allocation concealment (79%) and blinding (79%) and lower for sequence generation (59%) and incomplete outcome data (56%).

Conclusions: Most trial pairs found to be discrepant empirically were judged to be equally biased by assessors. We found moderate agreement between opinion and data-based evidence in pairs where assessors ranked one trial as more biased.

Keywords:

Meta-analysis, Systematic reviews, Randomised trials, Bias

What is new?

Key findings

- We found moderate agreement between opinion and data-based evidence in the rankings of pairs of randomised trials by bias severity, in pairs where assessors ranked one trial as more biased.
- Most trial pairs found to be discrepant empirically under a bias model fitted to meta-epidemiological data were judged to be equally biased by assessors.

What this adds to what was known

- Methods for bias adjustment in meta-analysis have been proposed by a number of authors and are usually informed by empirical evidence or elicited expert opinion on bias.
- The extent to which assessors' opinions on bias are similar to empirical estimates informed by meta-epidemiological research has not previously been evaluated.
- Bias adjustment can be informed by a combination of empirical evidence and opinion, with the aim of reducing uncertainty by using knowledge of the specific studies included in a meta-analysis.

What is the implication and what should change now

- Our finding that the majority of trial pairs were ranked as equally biased suggests that incorporating opinion on bias may not reduce uncertainty much, compared with using empirical distributions for bias alone.

INTRODUCTION

A meta-analysis of the results from relevant randomised trials is often regarded as the best evidence evaluating the effectiveness of a healthcare intervention (1). Meta-analysis results summarise the findings from multiple studies and are more precise and usually more influential than results from a single trial. Their findings inform public health policy decisions made by organisations such as the National Institute for Health and Care Excellence (NICE), as well as healthcare decisions made by individual patients, doctors and institutions. Randomised trials vary in methodological quality, and flaws in trial conduct can lead to biased estimation of the intervention effect (2). If a meta-analysis makes no allowance for methodological flaws, there is a danger that the results could be biased and more precise than they should be (3), which can lead to inappropriate healthcare decisions.

Randomised trials should employ rigorous methods that minimise the risk of bias and preserve comparability of the intervention groups. For example, concealment of randomised allocation ensures that the order of assignments to intervention groups cannot be predicted in advance and thereby removes the influence of patient characteristics on the probability of assignment to a group. Blinding of participants and caregivers to randomised allocation prevents differences in patient management between groups and blinding of outcome assessors (including participants when outcomes are reported by them) prevents knowledge of allocation influencing outcome measurement. Inadequacies in allocation concealment and blinding have been found to be associated with exaggeration of intervention effects (4-8). Meta-analyses often include trials that vary in methodological adequacy with respect to these characteristics and others.

Assessing the risk of bias in included studies is a mandatory step in a systematic review (9, 10) but there is no established method for combining bias assessments with a meta-analysis to guide interpretation of the effect of an intervention. The majority of systematic reviews do not incorporate bias assessments into the statistical analysis (11). In those which do incorporate bias assessments, the most common approach is to perform a sensitivity analysis excluding high risk studies, following a primary analysis including all evidence. This is problematic because it requires researchers to categorise available trials as either “good” and eligible for inclusion or “bad” and to be excluded. In

many meta-analyses, a criterion to dichotomise trials as good or bad is not easily chosen and, if few trials remain eligible for inclusion, precision could be greatly reduced. For example, 43% of trials were judged to be at high risk of bias for at least one domain of the Cochrane Risk of Bias tool (12), so exclusion on this basis could almost halve the number of trials included. Under this approach to addressing biases, discarded trials are regarded as providing no useful information at all, while included trials are implicitly assumed to be unaffected by within-trial biases. Most meta-analyses include trials which lie somewhere between these two extremes. Although sensitivity analyses based on risk of bias are often reported, decision making will usually be based on a single summary result, and it would therefore be desirable for the primary meta-analysis to incorporate adjustment for within-trial biases. Adjusting a meta-analysis for biases that are present in included trials is often considered controversial. However, the conventional approach of making no adjustment to the results even when potential causes of bias are present in a trial is equivalent to assigning an extremely strong opinion to the assumption that the bias is equal to zero.

Methods for bias adjustment in meta-analysis have been proposed by a number of authors, allowing the influence of evidence from less rigorous trials to be reduced in the combined analysis (3, 13-17). Although the potential causes of bias are often known, the impact of bias affecting each trial is unknown. Distributions describing the expected level of within-trial bias and the uncertainty about the bias are constructed from external evidence, which is typically in the form of expert opinion or relevant empirical data. Empirical evidence on biases affecting randomised trials is available from meta-epidemiological studies which analyse large numbers of meta-analyses to examine the association between trial design characteristics and trial results (18). Meta-epidemiological research has provided evidence on the biases associated with flaws in sequence generation, allocation concealment, blinding and incomplete outcome data (4-6, 19-21). Welton et al. (17) proposed a method which uses generic empirical evidence on the magnitude of biases, obtained from meta-epidemiological studies based on collections of meta-analyses. Turner et al. (16) proposed a method which uses elicited expert opinion on the likely magnitude of biases, informed by detailed assessment

of the trials in the meta-analysis. The extent to which assessors' opinions on bias are similar to empirical estimates informed by meta-epidemiological research has not previously been evaluated.

In some instances, it would be desirable for bias adjustment in meta-analysis to be informed by a combination of empirical evidence on bias and opinion. For example, available meta-epidemiological evidence may be considered only partially relevant to a specific meta-analysis because of a difference in population or intervention settings, and expert opinion could be used to adjust the data-based distribution for bias to the target setting. If relying on meta-epidemiological evidence alone, the predicted distribution for within-trial bias is often very imprecise, because it allows for variability in bias across the collection of meta-analyses. By using opinion informed by knowledge of the studies included in a meta-analysis, it is likely that this uncertainty can be reduced. Using a combination of data-based evidence and opinion for the reasons described above would be considered more valid if these approaches were known to produce similar estimates for bias.

In this research, we obtain opinions on the bias associated with four domains, using meta-analyses sampled from a meta-epidemiological study. Our aims were to examine agreement among experts and subsequently to explore agreement between empirical data-based and opinion-based approaches to assessing bias.

METHODS

Outline of our approach

The approach to adjusting for biases based on empirical evidence involves fitting a hierarchical model to the data from trials included in each of a collection of meta-analyses (17). For our investigations we used data from the Risk of Bias in Evidence Synthesis (ROBES) study (6). Within each meta-analysis extracted from the ROBES database, we selected the two trials with the highest and lowest model-based bias estimates, and then elicited opinion on which trial was judged to be more biased. We examined agreement between model-based and opinion-based estimates of bias within selected pairs of trials.

ROBES study

The ROBES database consists of meta-analyses extracted from the April 2011 issue of the *Cochrane Database of Systematic Reviews*, in which Cochrane review authors had implemented the 'risk of bias' tool to assess potential biases in included trials (22). The ROBES study (6) included 228 meta-analyses in total, from Cochrane reviews that reported information on all five recommended risk of bias domains: sequence generation, allocation concealment, blinding, incomplete outcome data and selective outcome reporting. Review authors had recorded whether there was a low, high or unclear risk of bias in each bias domain, together with comments or quotes from the trial publication to justify each judgement. Meta-analyses were excluded if they included fewer than five trials or if a summary estimate was not reported in the review (for example, because pooling was considered inappropriate). One or more binary outcome meta-analyses (with sets of included trials that were unique to each meta-analysis) from each eligible review were included in the ROBES database; primary outcomes were chosen where possible (6).

Selection of pairs of trials within meta-analyses

For each meta-analysis, we selected a pair of trials with the highest and lowest model-based bias estimates, representing the least and the most biased trials among those included in the meta-analysis, for each of four bias domains: allocation concealment; sequence generation; blinding and incomplete outcome data. These pairs were selected in order to present them to expert assessors, asking them which trial of each pair they judged to be at the greatest risk of bias in each domain examined. The process of selecting pairs of trials is described in detail below.

For each bias domain in turn, we first sampled 30 meta-analyses from the ROBES study. Meta-analyses included in the ROBES study were sampled from the *Cochrane Database of Systematic Reviews* in April 2011. Meta-analyses were sampled from the set of meta-analyses including at least one trial judged to be at low risk of bias and at least two trials judged to be at high or unclear risk of bias. A trial at low risk was needed as a comparator, to enable bias estimates to be obtained for trials with high or unclear risk judgements; at least two of the latter were required in order that the two with highest and lowest bias estimates could be selected. For example, when sampling meta-analyses to examine the bias associated with allocation concealment, we sampled 30 meta-analyses including at

least one trial assessed by review authors to have adequate allocation concealment and at least two trials assessed to have inadequate or unclear allocation concealment. To ensure that different outcome types were represented, each set of 30 meta-analyses comprised randomly selected samples of 15 eligible meta-analyses with outcomes judged to be objective or semi-objective ("objectively ascertained but potentially influenced by judgement") in the ROBES study and 15 eligible meta-analyses with outcomes judged to be subjective or of mixed types within the meta-analysis (6). Choice of sample size of 30 meta-analyses per bias domain is justified in the Appendix.

For each bias domain in turn, we fitted the bias model proposed by Welton et al to all meta-analyses in the ROBES database and obtained estimates (together with uncertainty) for the trial-specific biases within the 30 sampled meta-analyses. The binary outcome data r_{mia}, n_{mia} (representing number of events and total number of subjects) from each trial arm *a* of trial *i* within meta-analysis *m* were assumed to have a binomial likelihood, $r_{mia} \sim Bin\left(p_{mia}, n_{mia}\right)$. The following hierarchical bias model includes effects of trial-specific biases β_{mi} associated with a known trial characteristic Z_{mi}, and allows for within-meta-analysis bias variation κ^2 and between-meta-analysis bias variation ϕ^2 (17). Treatment effects δ_{mi} are assumed random across trials within meta-analyses, with separate between-trial heterogeneity variances τ_m^2. The values of δ_{mi} and τ_m^2 were assumed to be unrelated across meta-analyses.

$$\begin{aligned} \text{logit}\left(p_{mia}\right) &= \mu_{mi} + X_{mi}\left(\delta_{mi} + \beta_{mi} Z_{mi}\right) \\ \delta_{mi} &\sim N\left(d_m, \tau_m^2\right) \\ \beta_{mi} &\sim N\left(b_m, \kappa^2\right) \\ b_m &\sim N\left(b_0, \phi^2\right) \end{aligned} \qquad (1)$$

Posterior mean values of the β_{mi} were used as bias estimates, and viewed as model-based assessments for the extent of bias in particular trials. These are shrinkage estimates of bias, based on borrowing information across the meta-analyses in the ROBES database.

Next, for each bias domain in turn and within each sampled meta-analysis, we selected the pair of trials with the highest and lowest bias estimates, among the trials with a judgement of high or unclear risk of bias. Selected pairs of trials from each of the sampled meta-analyses formed our study data set in which empirical data-based and opinion-based approaches to assessing bias were compared.

Elicitation of opinion on bias

Every trial in each pair was summarised by a description of the trial participants, interventions, outcomes and methods (together with additional notes, if available), extracted from the study characteristics tables reported by Cochrane reviewers. Trial sample sizes were added to each trial design summary, but no treatment effect estimates were provided. Support text for risk of bias judgements (without the actual judgements) was extracted from the Cochrane risk of bias tables for each trial and included in the summary information and checked against the original trial reports by the research team. If no support text was available in the risk of bias table or if it was incomplete, vague or not directly relevant to the given bias domain, it was extracted from trial reports by the research team.

We recruited six assessors (AH, DC, RM, BCR, HW, IB) with expertise in clinical research methodology and evidence-based medicine, by personal invitation. For each trial pair, assessors were given information packs (see example in Appendix) and asked to complete them independently. In total, each trial pair was assessed three times, by three out of six assessors. Trials within the pairs were labelled “trial A” and “trial B” at random. For each of four bias domains (sequence generation, allocation concealment, blinding, incomplete outcome data), the assessors were asked to choose between the following three judgements: “trial A is more biased”, “trial B is more biased”, or “trial A and trial B are equally biased”. We note that assessors were asked to make judgements for all four bias domains, without knowledge of the bias domain for which the trial pair had been selected. In addition, assessors were asked to choose between the same three judgements with respect to overall risk of bias. Alongside each judgement, assessors were asked to provide a rating from 1 to 5 for their confidence in that judgement, where 1 represents “not at all confident” and 5 represents “very confident”. The assessors attended a one-day meeting to carry out their rankings and were asked not

to discuss their judgements with other assessors; several assessors required more time and completed the work later on.

Data analysis

We examined agreement in the trial pair rankings (ordering of trials A and B with respect to extent of bias) among the bias assessors, using unweighted kappa statistics and 95% confidence intervals. Analyses were performed for each bias domain separately and then for all bias domains combined, using rankings from all trial pairs in the study data set.

We assessed agreement between the trial pair rankings produced by assessors and the ranking based on estimated biases from the bias model. We reported the proportion of trial pairs in which assessors chose one trial as more biased (rather than saying they were equally biased). Of the judgements in which one trial was believed to be more biased than the other, we calculated the proportion in which assessor opinion agreed with the model-based ranking of the trials. Analyses were performed for each bias domain separately, using the rankings from the subset of 30 meta-analyses sampled for that bias domain.

Next, we conducted exploratory multinomial logistic regression analyses to examine the association between assessor opinions and model-based differences in bias estimates between the trials in each pair. We used regression to explore whether agreement between assessor ranking and model-based ranking was associated with the magnitude of the difference in estimated biases for each trial pair. For each combination of trial pair (i) and assessor (j), there are three possible outcomes: disagreement between the assessor and model-based rankings, agreement between the assessor and model-based rankings, or assessors ranking trials as equally biased. Disagreement between the assessor and model-based ranking was treated as the baseline category (k=0) for the response variable, and a multinomial logistic regression model was created to estimate the odds ratio for each of the two alternative categories: assessors ranking trials as equally biased (k=1), and assessors agreeing with the model-based ranking (k=2). As a single covariate in the model we included the magnitude of difference in bias estimates in the trial pair. The multinomial logistic regression model was:

$$logit\left(\pi_{ijk}\right) = \alpha_k + \beta_k x_i + u_{ik} + \gamma_{jk}, \quad (2)$$

where π_{ijk} represents the probability of outcome category *k* for assessor j in trial pair *i*, and x_i is the model-based difference in bias estimates (calculated as the difference between the most extreme and least extreme bias values). To allow for similarity in judgements on the same trial pair (or equivalently, variation between trial pairs), we included a random intercept u_i for each of the 30 trial pairs. We also included a fixed effect γ_j for each of the six different assessors. We focus on the regression coefficient β_2 of the model-based difference in bias estimates. A positive value for this coefficient indicates that, on average, assessor agreement with model-based rankings is associated with the magnitude of the estimated difference in bias from the model.

All regression models were fitted using MCMC methods within WinBUGS (23) (see Appendix).

RESULTS

Descriptive analyses

Our data set consisted of 101 trial pairs in total because there was some overlap between the sets of 30 meta-analyses sampled for each of the four bias domains. Table 1 summarises the types of interventions and outcomes evaluated in the sampled meta-analyses. The majority (64%) of sampled meta-analyses corresponded to pharmacological vs. placebo/control comparisons, while 25% were non-pharmacological vs. control comparisons, and the remainder represented comparisons of two active treatments. Objective outcomes were evaluated in 36% of sampled meta-analyses overall, 16% evaluated semi-objective ("objectively ascertained but potentially influenced by judgement") outcomes and 46% evaluated subjective outcomes. The median number of trials included in the meta-analyses was 13 (inter-quartile range 9 to 24). Meta-analysis characteristics were fairly similar across the meta-analysis samples selected for each bias domain (Table 1).

Table 2 shows the frequencies of risk of bias profiles (combinations of risk of bias judgements for the four bias domains, reported by Cochrane authors) among the trials selected as having the lowest or highest bias estimates within meta-analyses. Of 202 trials, 120 (59%) had judgements of high or

unclear risk of bias for three or four bias domains and no trials had low risk of bias judgements for all domains. Differences within trial pairs are summarised in Table 3. Risk of bias judgements differ within pairs for only one bias domain or no bias domains in 59/101 trial pairs, and differ for all four bias domains in only 4/101 pairs.

Table 4 describes the extent of agreement among the bias assessors when judging which trial of each pair they believed to be more biased, showing the estimated kappa statistics in the rankings of the three assessors. There was fair to moderate agreement among the rankings. For sequence generation, the percentage of pairs in which all three assessments agreed was 50% and the kappa statistic was 0.43 (95% CI: 0.37 to 0.50). For allocation concealment, the percentage in which all three assessors were in agreement was 56% and the kappa statistic was 0.46 (95% CI: 0.40 to 0.52). There was moderate agreement among rankings for blinding: the percentage agreement across all three assessors was 60% and kappa was estimated as 0.45 (95% CI: 0.39 to 0.51). There was less agreement among assessors for incomplete outcome data: the percentage in which all three assessors agreed was 31% and the kappa statistic was 0.21 (95% CI: 0.14 to 0.27). For overall risk of bias, the percentage of trial pairs in which all three assessors agreed was 32% and the kappa statistic was 0.26 (95% CI: 0.19 to 0.32).

The assessors specified a confidence level of 1 (not at all confident) to 5 (very confident) about their opinion. We summarize the confidence levels in Figure 1. Assessor confidence levels were comparable for sequence generation, allocation concealment and blinding. For each of these bias domains, the median confidence level across all trial pairs and all assessors was 3 (inter-quartile range (IQR): 2 to 4). Confidence levels tended to be lower for incomplete outcome data and for overall bias (median 2, IQR: 1 to 3 for each). Confidence levels were no higher when examined only in the bias domain for which the trial pair had been selected.

For each bias domain, 30 trial pairs were ranked by each of three assessors, resulting in 90 assessor opinions. For sequence generation, 36 (40%) of the 90 assessor opinions ranked one trial as more biased than the other (Table 2). For allocation concealment, blinding and incomplete outcome data respectively, 14 (16%), 24 (27%) and 41 (46%) of opinions ranked one trial as more biased. Table 5

reports the proportion of assessor opinions that agreed with the model-based ranking of trial pairs. Among the assessor opinions which judged one trial as more biased (rather than trials equally biased), the proportion that agreed with the ranking based on the bias model was high for allocation concealment (79%) and blinding (79%). For sequence generation and assessment of incomplete outcome data, agreement was lower at 59% and 56% respectively (i.e. not much better than chance).

Regression analyses

In the exploratory multinomial logistic regression analyses, we focus on the regression coefficient β_2 of the model-based difference in fitted bias (Table 6). Although this was estimated as positive for allocation concealment and incomplete outcome data, the 95% credible intervals were very wide and contained the null value, representing no association between the magnitude of difference in model-based bias estimates and agreement between assessor and model-based rankings. For sequence generation and blinding, the regression coefficient was estimated as negative, again with very wide 95% credible intervals containing the null value. Similarly, we cannot conclude whether smaller differences in model-based bias estimates were associated with assessors ranking trials as equally biased. There is insufficient information in the data for us to be able to draw any conclusions from the results (Table 6): all intervals for model parameters were wide and close to the ranges of the assigned prior distributions.

DISCUSSION

Agreement between opinion-based and model-based rankings of bias magnitude was high for sequence generation and allocation concealment and moderate for blinding and incomplete outcome data, among trial pairs in which assessors ranked one trial as more biased. However, in the majority of trial pairs, assessors ranked trials as equally biased, although the two trials had been selected on the basis of having high and low bias estimates (within a given meta-analysis) under the bias model fitted. In these trial pairs, detailed trial descriptions did not lead assessors to judge the bias as higher in one trial than another. There was fair to moderate agreement in rankings across bias assessors. In exploratory regression analyses, uncertainty was too high for us to draw conclusions about

associations between the magnitude of difference in model-based bias estimates and assessors agreeing with model-based rankings or assessors ranking trials as equally biased.

Published methods for bias adjustment in meta-analysis suggest making use of either empirical data-based evidence on biases or opinion on biases (3, 13, 15-17), but no previous comparison has been made between data-based distributions and assessors' opinions on bias. Access to a large collection of meta-analyses for which review authors have reported risk of bias judgements and supporting information has enabled us to carry out a comparative study. We note that the empirical data-based distributions for bias were themselves informed indirectly by opinion, since they were derived from a hierarchical model fitted to trial data within meta-analyses, in which judgement about each trial's risk of bias was used as a covariate. The model-based rankings rely on the appropriateness of the assumed model for the data and also on the risk of bias judgements reported by Cochrane reviewers. Reviewers follow risk of bias protocols that aim to maximise reproducibility. It would not be possible to adjust a meta-analysis for trial-specific biases without incorporating some form of subjective judgement. Formal validation methods are not available for bias assessments, since the true extent of bias in a given trial is unknown, but agreement between independent bias assessments would increase our confidence in them.

Since the actual magnitude of bias affecting the trial pairs selected from the sampled meta-analyses remains unknown, it is not possible to evaluate whether the data-based or opinion-based rankings are closer to the truth. Assessors indicated that their confidence in their own opinions on the rankings of trials within pairs was moderate or low. In our study, assessors were asked to carry out a large number of rankings during one day (though several assessors required more time and completed the work later on); the high workload may have affected their performance. When assessors are asked to provide opinions on biases affecting studies in a single meta-analysis, the number of studies assessed would typically be much smaller. We observed less agreement among assessors for incomplete outcome data than for the other bias domains. This may be related to the greater complexity of the bias in this domain, which depends on several factors, including the amount and distribution of missing data across intervention groups, the likely difference in outcome between missing and non-missing

participants, and how the problem has been addressed in reported analyses (24). We aimed to assess agreement among assessors pragmatically, so we did not attempt to increase inter-observer agreement before carrying out the elicitation exercise.

In this work, opinions about biases were based on summary information about trials, informed primarily by the study characteristics and risk of bias tables reported by Cochrane reviewers and supplemented by additional information extracted from the trial reports by the research team. Assessors reported some difficulties in assessing bias on the basis of summary information and commented that for certain trials they would have liked access to the original trial publications. When eliciting opinions about within-trial biases, it might therefore be preferable to provide full publications, as Turner et al. did in their opinion-based method for bias adjustment (16), although this introduces some risk that assessments of bias are influenced by knowledge of the trial results unless all results are removed. Using all available sources of information (e.g. publication, statistical analysis plan, protocol, trial registration records etc.) is generally encouraged for assessing risk of bias in RCTs included in systematic reviews (25), to improve confidence in assessment. We were surprised that the majority of trial pairs were ranked as equally biased and we suspect that the lack of detailed trial information contributed to this. We expect that differentiation between trials was reduced also by requesting categorical judgements for each trial pair rather than continuous judgements of bias (using a visual analogue scale, for example) for each individual trial. Trials judged to be at high or unclear risk of bias were grouped together in the hierarchical model used to estimate bias. Research has suggested that many trials judged to be at unclear risk of bias for sequence generation and allocation concealment could be reclassified as low risk if information outside the trial publications was obtained (26). Misclassification of risk of bias judgements may have reduced or increased the differences within some of the selected trial pairs.

Risk of bias judgements are increasingly published for trials included in Cochrane reviews. It is desirable to incorporate these judgements about suspected biases into the statistical analyses performed and interpretation of the review findings (10, 11). The Cochrane database could in time provide extensive evidence on the degree of bias associated with combinations of risk of bias

judgements for different domains. In a separate paper, we have explored methods for quantifying bias by using empirical distributions for the bias affecting trials with a specific set of risk of bias judgements, in combination with expert opinion (27). However, our finding in this paper that the majority of trial pairs were ranked as equally biased suggests that incorporating opinion on bias may not reduce uncertainty much, compared with using empirical distributions alone.

We found moderate agreement between opinion and data-based evidence in the rankings of trial pairs by bias severity, in pairs where assessors ranked one trial as more biased. This finding provides some support for approaches combining data-based evidence with opinion on bias. However, trials were ranked as equally biased in the majority of trial pairs, indicating that trial summaries did not provide sufficient information to reach a ranking judgement.

Acknowledgements

This project was funded by the UK Medical Research Council (MRC) grant (MR/K014587/1). RMT and KMR were supported by the MRC programme grant (U105260558). RMT was also supported by the MRC grant MC_UU_12023/21 and KMR also by the MRC grant MC_UU_00002/5. HEJ was supported by a MRC career development award in biostatistics (MR/M014533/1). JPTH is an NIHR Senior Investigator (NF-SI-0617-10145), a member of the MRC Integrative Epidemiology Unit at the University of Bristol, and is supported by the NIHR Applied Research Collaboration West (ARC West) at University Hospitals Bristol NHS Foundation Trust and the NIHR Bristol Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. JS was supported by the MRC fellowship (G0701659/1) and the National Institute for Health Research (NIHR) ARC West at University Hospitals Bristol NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the MRC, the National Health Service, the NIHR or the Department of Health and Social Care.

Declaration of interests

Isabelle Boutron is co-convenor of the Cochrane Bias Methods Group.

References

1. Egger M, Davey Smith G, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001.
2. Gluud LL. Bias in clinical intervention research. American Journal of Epidemiology. 2006;163:493-501.
3. Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). Journal of the Royal Statistical Society, Series A. 2005;168:267-306.

4. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal American Medical Association. 1995;273:408-12.
5. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Annals of Internal Medicine. 2012;157:429-38.
6. Savovic J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association Between Risk-of-Bias Assessments and Results of Randomised Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. Am J Epidemiol. 2018;187(5):1113-22.
7. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272-83.
8. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119.
9. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
10. Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of Cochrane Intervention Reviews. Version 1.06 ed. London: Cochrane; 2018.
11. Hopewell S, Boutron I, Altman DG, Ravaud P. Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. BMJ Open. 2013;3(8):e003342.
12. Yordanov Y, Dechartres A, Porcher R, Boutron I, Altman DG, Ravaud P. Avoidable waste of research related to inadequate methods in clinical trials. BMJ. 2015;350:h809.
13. Eddy DM, Hasselblad V, Schachter R. Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence. San Diego, CA: Academic Press; 1992.
14. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Statistics in Medicine. 2003;22:3687-709.
15. Wolpert RL, Mengersen KL. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. Statistical Science. 2004;19:450-71.
16. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society, Series A. 2009;172:21-47.
17. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society, Series A. 2009;172(1):119-36.
18. Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. BMJ. 1997;315(7109):617-9.
19. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? The Lancet. 1998;352:609-13.
20. Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomised controlled trials. Journal American Medical Association. 2002;287:2973-82.
21. Savovic J, Jones HE, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technology Assessment. 2012;16:1-82.
22. Higgins JPT, Altman DG. Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Chichester: John Wiley & Sons; 2008.

23. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2000;10(4):325-37.
24. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
25. A revised tool to assess risk of bias in randomised trials (RoB 2) [Available from: https://www.riskofbias.info/welcome/rob-2-0-tool.
26. Vale CL, Tierney JF, Burdett S. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. BMJ. 2013;346:f1798.
27. Rhodes KM, Savovic J, Elbers R, Jones HE, Higgins JPT, Sterne JAC, et al. Adjusting trial results for biases in meta-analysis: combining data-based evidence on bias with detailed trial assessment. Journal of the Royal Statistical Society, Series A. 2020;183:193-209.

Table 1 Characteristics of meta-analyses sampled from the ROBES data set, for each bias domain and overall.

Characteristics of meta-analyses sampled	**Bias domain**				
	Sequence generation (n=30)	Allocation concealment (n=30)	Blinding (n=30)	Incomplete outcome data (n=30)	Overall (n=101)
Type of intervention comparison					
Pharmacological vs. Placebo/control	17 (57%)	21 (70%)	19 (63%)	20 (67%)	65 (64%)
Pharmacological vs. Pharmacological	5 (17%)	1 (3%)	1 (3%)	2 (7%)	8 (8%)
Non-pharmacological vs. Placebo/control	8 (27%)	7 (23%)	8 (27%)	7 (23%)	25 (25%)
Non-pharmacological vs. Non-pharmacological	0 (0%)	1 (3%)	2 (7%)	1 (3%)	3 (3%)
Type of outcome measure					
Objective	11 (37%)	11 (37%)	10 (33%)	11 (37%)	36 (36%)
Semi-objective	5 (17%)	4 (13%)	5 (17%)	3 (10%)	16 (16%)
Subjective	13 (43%)	14 (47%)	14 (47%)	15 (50%)	46 (46%)
Mixed types within the meta-analysis	1 (3%)	1 (3%)	1 (3%)	1 (3%)	3 (3%)
Number of trials: median (interquartile range)	13.5 (10 to 20)	13.5 (9 to 24)	12 (8 to 18)	15 (9 to 24)	13 (9 to 24)

Table 2 Frequencies of risk of bias profiles (from Cochrane reviews) in trials selected from sampled meta-analyses

Bias domain				Frequency (%) (n=202)
SG	**AC**	**B**	**IOD**	
				0 (0%)
				0 (0%)
				6 (3%)
				13 (6%)
				7 (3%)
				20 (10%)
				7 (3%)
				3 (1%)
				8 (4%)
				7 (3%)
				11 (5%)
				34 (17%)
				20 (10%)
				7 (3%)
				5 (2%)
				54 (27%)

SG: sequence generation; AC: allocation concealment; B: blinding; IOD: incomplete outcome data

 High/Unclear risk of bias

 Low risk of bias

Table 3 Differences in risk of bias profiles (from Cochrane reviews) within trial pairs

	Frequency (%) (n=101)
High/unclear/low judgements match for all bias domains	23 (23%)
Difference in judgements for one bias domain	36 (36%)
Differences in judgements for two bias domains	27 (27%)
Differences in judgements for three bias domains	11 (11%)
Differences in judgements for four bias domains	4 (4%)

Table 4 Kappa statistics with 95% confidence intervals for assessing agreement in rankings among the three bias assessors.

Bias domain	**Trial pairs**	**Un-weighted kappa (95% CI)**	**Interpretation**	**% trial pairs with three assessments in agreement**
Sequence generation	All 101	0.43 (0.37 to 0.50)	Moderate agreement	50/101 (50%)
Allocation concealment	All 101	0.46 (0.40 to 0.52)	Moderate agreement	57/101 (56%)
Blinding	100[1]	0.45 (0.39 to 0.51)	Moderate agreement	60/100 (60%)
Incomplete outcome data	99[1]	0.21 (0.14 to 0.27)	Fair agreement	31/99 (31%)
Overall	97[1]	0.26 (0.19 to 0.32)	Fair agreement	31/97 (32%)

[1] Missing expert opinions

Table 5 Frequency of assessor opinions ranking one trial as more biased (rather than choosing trials equally biased). Of those that chose one trial as more biased, we report the proportion that agreed with the fitted model of Welton *et al*.

Bias domain	**How often did the assessors choose one trial as more biased (rather than equally biased)?**	**Of those that chose one trial as more biased, what proportion agreed with the model?**
Sequence generation	36/90 (40%)	23/36 (59%)
Allocation concealment	14/90 (16%)	11/14 (79%)
Blinding	24/90 (27%)	19/24 (79%)
Incomplete outcome data	41/90 (46%)	23/41 (56%)

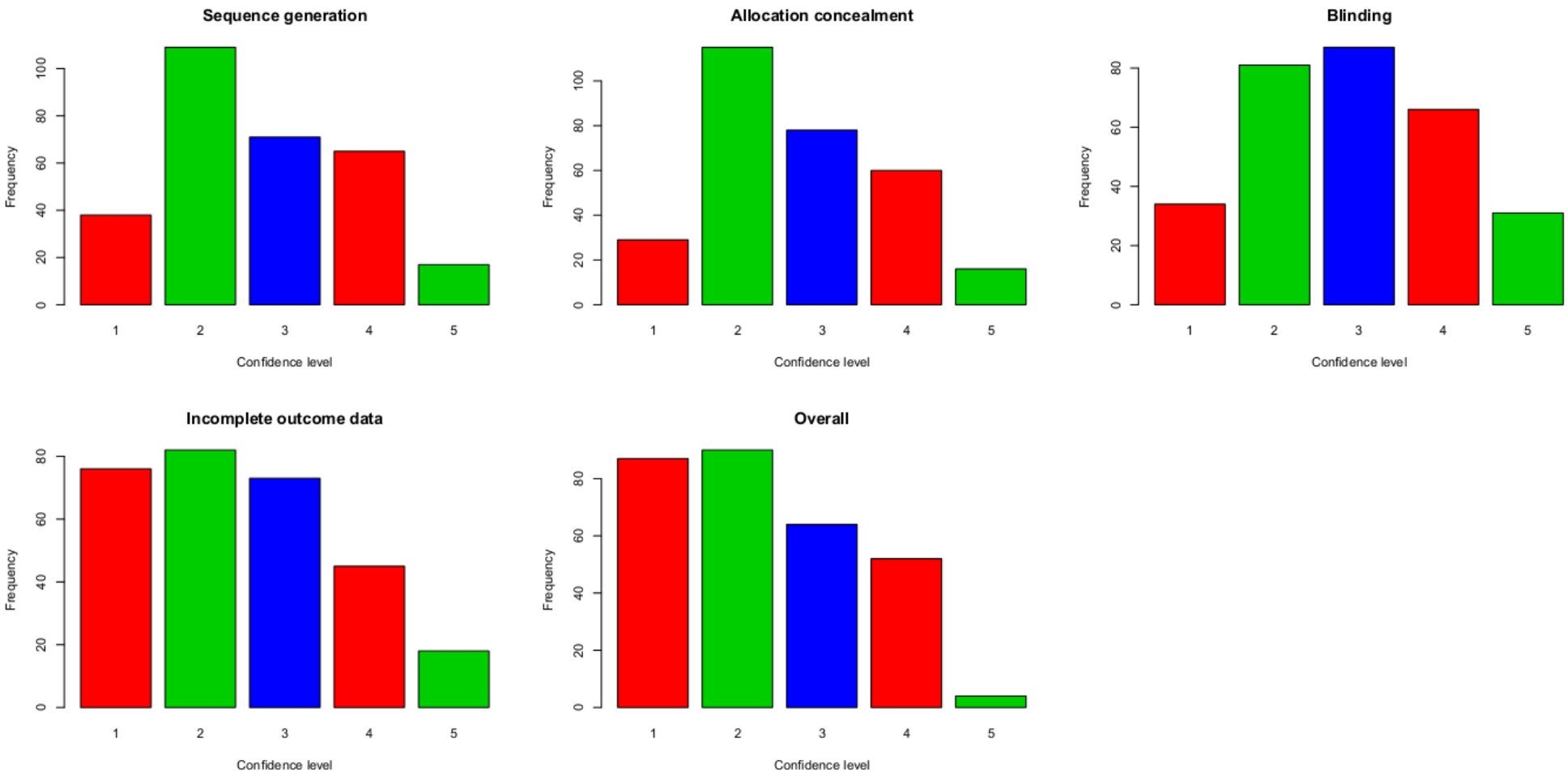
Table 6 Results from the exploratory multinomial regression to examine the association between assessor opinion and model-based difference in bias estimates: central parameter estimates (95% credible intervals).

Outcome		Sequence generation	Allocation concealment	Blinding	Incomplete outcome data
Assessor and model-based rankings agree	Model-based difference in bias estimates[1] (β_2)	-0.07 (-6.25 to 6.03)	0.04 (-6.20 to 6.23)	-0.08 (-6.32 to 6.08)	0.42 (-5.80 to 6.60)
	Assessor effects $(\alpha_2 + \gamma_{j2})$				
	1	-0.77 (-2.95 to 1.06)	-1.21 (-4.67 to 1.79)	0.08 (-4.07 to 4.11)	-1.47 (-4.44 to 1.03)
	2	N/A	-2.50 (-7.61 to 2.08)	1.98 (-0.04 to 4.38)	0.10 (-1.90 to 1.86)
	3	-0.52 (-3.88 to 2.80)	1.24 (-1.08 to 3.84)	-1.20 (-4.66 to 1.95)	-0.22 (-2.42 to 1.74)
	4	0.27 (-2.14 to 2.61)	-0.35 (-4.39 to 3.52)	0.46 (-1.59 to 2.40)	-0.02 (-3.22 to 3.05)
	5	1.32 (-1.66 to 4.50)	-0.86 (-4.29 to 2.18)	-1.76 (-6.94 to 2.64)	0.37 (-1.70 to 2.39)
	6	1.59 (-0.93 to 4.34)	1.05 (-2.22 to 4.39)	0.50 (-2.82 to 3.93)	-0.13 (-2.35 to 1.99)
	Between-trial-pair standard deviation	3.31 (1.43 to 4.88)	2.43 (0.06 to 4.70)	2.17 (0.36 to 4.47)	2.35 (0.82 to 4.52)
Assessor and model-based rankings disagree	Baseline outcome				
Trials equally biased	Model-based difference in bias estimates[2] (β_1)	-0.23 (-6.47 to 6.05)	-0.37 (-6.57 to 5.86)	-0.48 (-6.61 to 5.64)	0.14 (-6.07 to 6.23)
	Assessor effects $(\alpha_1 + \gamma_{j1})$				
	1	0.80 (-1.12 to 2.71)	3.80 (1.82 to 6.43)	3.02 (-0.27 to 7.03)	-0.59 (-3.00 to 1.51)
	2	N/A	5.14 (1.79 to 9.52)	2.80 (0.92 to 5.13)	1.11 (-0.45 to 2.76)
	3	1.10 (-2.47 to 4.61)	2.59 (0.51 to 5.00)	2.67 (0.58 to 5.37)	2.04 (0.53 to 3.90)
	4	2.62 (0.52 to 4.99)	4.67 (1.86 to 8.50)	1.10 (-0.84 to 2.89)	1.99 (-0.36 to 4.76)
	5	1.86 (-0.89 to 4.89)	2.90 (0.76 to 5.56)	5.07 (2.20 to 9.22)	1.08 (-0.67 to 2.89)
	6	3.94 (1.43 to 6.85)	4.32 (1.78 to 7.57)	4.76 (2.20 to 8.10)	0.94 (-0.91 to 2.90)
	Between-trial-pair standard deviation	3.99 (2.21 to 4.95)	2.03 (0.41 to 4.49)	1.84 (0.37 to 4.20)	1.94 (0.59 to 3.90)

[1] A positive value for β_2 indicates that, on average, greater differences in estimated bias within trial pairs are associated with assessor rankings agreeing with the model-based rankings.

[2] A positive value for β_1 indicates that, on average, greater differences in estimated bias within trial pairs are associated with assessors ranking trials as equally biased.

Figure 1 The confidence of assessors in their opinions on each bias domain and overall bias, where 5 represents "very confident" and 1 represents "not at all confident".



Appendix

Sample size justification

Choice of sample size of 30 meta-analyses per bias domain was based on considering precision of estimation of the proportion of pairs for which model-based and opinion-based orderings of biases agree. Overlap between the sets of 30 meta-analyses sampled for different bias domains was expected, meaning that the total number of sampled meta-analyses was likely to be less than 120. In a sample of 90 trial pairs (assuming 25% reduction), the standard error for the proportion of pairs with agreement would be less than 0.05 (maximum value 0.049 for an observed proportion of 0.5), assuming a high between-assessor correlation of 0.8. In each bias domain sample of 30 trial pairs, the standard error would be less than 0.1.

Model fitting

All regression models were fitted using MCMC methods within WinBUGS (23). We declared vague normal(0,10) priors for unknown regression coefficients and a uniform(0,5) prior for all standard deviations. Results were based on 100,000 iterations following a burn-in period of 10,000 iterations which was sufficient to achieve convergence. Convergence was checked using the Brooks-Gelman-Rubin statistic, as implemented in WinBUGS (23), with three chains starting from widely dispersed initial values.

Example information pack for a trial pair

Instructions: The two trials laid out below are taken from the same meta-analysis. Please compare these two trials with respect to the magnitude of bias due to inadequacies in sequence generation, allocation concealment, blinding and incomplete outcome data.

Review details

Review ID: 736 CD number: CD005496

Review title: Probiotics for prevention of necrotizing enterocolitis in preterm infants

Participants/populations: Preterm infants < 37 weeks and/or birth weight < 2500 g.

Interventions: Enteral administration of any live microbial supplement (probiotics) at any dose for more than seven days compared to placebo or no treatment.

Meta-Analysis details

Meta-analysis ID: 17134

Comparison details: Probiotics vs. Placebo or no treatment

Experimental intervention: Probiotics

Comparator intervention: Placebo or no treatment

Outcome: Severe necrotising enterocolitis (stage II-III)

TRIAL A

Trial details Trial ID: 21227 Trial name: Samanta 2009

Methods	Prospective randomised double-blind control trial Method of generating randomisation sequence: Can't tell Allocation concealment: Can't tell Blinding of intervention: Can't tell Blinding of outcome measurement: Can't tell Complete follow-up: Yes
Sample size	61
Participants	Gestational age <32 weeks and VLBW infants (<1500 g) started feed enterally and survived beyond 48 h of life Demographic data: Probiotics Group N=91, gestational age 30.12 (weeks) (1.63), birth weight 1172 (143) Control Group N=95, gestational age 30.14 (weeks) (1.59), birth weight 1210 (143)
Interventions	The probiotic group received a probiotic mixture (Bifidobacteria infantis, Bifidobacteria bifidum, Bifidobacteria longum and Lactobacillus acidophilus, each 2.5 billion CFU) with expressed breast milk twice daily, the dosage being 125 g kg -1 till discharge. The control group was fed with breast milk only.
Outcomes	Feed tolerance in terms of days required to reach full enteral feeding Length of hospital stay NEC Sepsis Death due to NEC or sepsis
Notes	Neonatal Care Unit of Medical College and Hospital, Kolkata, India Period of study: October 2007 - March 2008 Published: 2009 Source of Funding: not specified in paper

Risk of bias table Trial ID: 21227 Trial name: Samanta 2009

Bias Domain	Description of what was done (based on study reports/papers)
Adequate sequence generation?	Quote: "infants were randomly assigned to two groups by random number table sequence" (p.129) No further information provided.
Allocation concealment?	Not stated whether allocation was concealed prior to assignment Paper states this is a prospective randomised double-blind control trial, but detail not provided.
Blinding?	Paper states this is a prospective randomised double-blind control trial, but detail not provided. Quote: "the probiotic-fortified group received a probiotic mixture…with expressed breast milk daily…The control group was fed with breast milk only." (p.129) Not stated whether mothers or personnel were aware of allocation. Not stated whether assessors were blinded to allocation
Incomplete outcome data addressed?	Analysis appears to be based on full numbers of participants Attrition and exclusions from analysis were not reported.
Free of selective reporting?	All 3 primary outcome measures reported (feed tolerance, length of hospital stay, morbidities).
Free of other bias?	Birth weight and gestational age were not significantly different between groups. No other statitistically significant demographic or clinical variables between groups. Adverse effects not reported No sample size calculations reported Inclusion/exclusion criteria reported

TRIAL B

Trial details Trial ID: 21228 Trial name: Sari 2010

Methods	Single Center Method of generating randomisation sequence: Sequential numbers generated at the computer center of the NICU Allocation concealment: Can't tell Blinding of intervention: Can't tell Blinding of outcome measurement: Yes Complete follow-up: Yes
Sample size	63
Participants	Gestational age <33 weeks or birth weight <1500 g Demographic data: Probiotics Group N=110, gestational age 29.5 (weeks) (2.4), birth weight 1231 (262) Control Group N=111, gestational age 29.7 (weeks) (2.4), birth weight 1278 (282)
Interventions	VLBW infants who survived to start enteral feeding were randomised The study group were given L. sporogenes with a dose of 350.000.000 colony forming units added to breast milk or formula once a day starting with first feed until discharge. The control group were fed without L. sporogenes supplementation.
Outcomes	Death or severe NEC NEC (stage 2, 3, = 2) Death (attributable to NEC, not attributable to NEC) Total parental nutrition Intraventricular hemorrhage, grade 3-4, Sepsis (culture proven, gram negative, gram positive, fungus) NICU stay Feeding (amount, full feeding, intolerance) Weight gain
Notes	Turkey Period of study: October 2008 and June 2009 Published: Unpublished Source of Funding: not specified in paper

Risk of bias table Trial ID: 21228 Trial name: Sari 2010

Bias Domain	Description of what was done (based on study reports/papers)
Adequate sequence generation?	Quote: "The infants were randomly assigned to one of two groups prospectively. Randomization was performed by using sequential numbers generated at the computer center of the NICU" (p.435)
Allocation concealment?	Quote: "The allocations were contained in opaque, sequentially numbered sealed envelopes" (p.435)
Blinding?	Blinding of intervention: Quote: "the only personnel who knew of the infants' group assignments were the investigators and those in the breast-milk team who were not involved in the care of the study infants" (p.435) Supplementation given to the experimental group did not alter the appearance of the milk or formula. Quote: "Fresh suspension of supplements were prepared by personnel in the breast-milk team who were not involved in the care of the infant and who followed instructions from the sealed envelope" (p.435) Blinding of outcome: Quote: "Whenever an infant was suspected to have NEC [outcome measure], the infant was evaluated by two senior-attending neonatologists who did not know the group assignment of the infant" (p.435)
Incomplete outcome data addressed?	Attrition and exclusions reported, no apparent imputation Analysis only of sample after attrition (11 lost in experimental group, 10 lost in control).
Free of selective reporting?	Primary and secondary outcome measures were reported Primary = death or stage >2NEC

	Secondary = culture proven sepsis without NEC, intraventricular hemorrhage, feeding intolerance, feeding amount per week, days to reach full enteral feeding, wight gain per week.
Free of other bias?	Clinical and demographic characteristics did not differ between groups, except for: Quote: " longer duration of umbilical venous catheterization in the probiotics group" (p.436) Possible effect of this discussed in discussion Inclusion/exclusion criteria reported Adverse effects were reported. Sample size calculations reported - number needed = 111 infants for each arm. Total sample = 110 experimental, 111 control Quote: "the required sample size was above the actual numbers attained in our study, which in turn make the study underpowered to detect small differences" (p.438)

EXPERT VERDICT

In which of these two trials would you expect inadequacies in each of the listed bias domains to cause greater bias towards overestimation of the effect estimate of treatment benefits for the experimental intervention?

Please consider each bias domain separately. For each bias domain, your decision should be explicit to that particular bias, but the whole risk of bias table may be taken into account (i.e. information provided under all other bias domain may influence your decision for any one particular domain).

Please tick only one verdict option per domain and indicate your level of confidence for each verdict on a scale of 1 to 5 (1 being not very confident at all and 5 being very confident).

Meta-analysis outcome: **Severe necrotising enterocolitis (stage II-III)**

	Your verdict on bias Tick only one of the three possible options for each bias domain (each row)			**Confidence level score** How confident are you about this verdict? Enter confidence score between 1 and 5 for each bias domain, where 1 = not at all confident; 5 = very confident.
Bias domain (tick only one box per row for each domain & overall bias)	Option 1: **Trial *A* is more biased**	Option 2: **Trial *B* is more biased**	Option3: **Trial A and Trial B are *equally* biased**	
Sequence generation				
Allocation concealment				
Blinding				
Incomplete outcome data				
Overall risk of bias for this outcome				

Expert Assessor notes *(optional)*